CHAPTER 1

INTRODUCTION

1.1 Introduction

Clustering is the unsupervised classification of patterns. It deals with finding a structure in a collection of unlabeled data. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification. The clustering of chemical compounds is a widely used technique in the field of chemo informatics for the selection of compounds for screening, the analysis of substructure searching, and the prediction of molecular properties and biological activities from structural information.

The Particle Swarm Optimization (PSO) is a population-based optimization method, was introduced by Eberhart and Kennedy [Eberhart and Kennedy, 1995]. It was originally developed for optimization in a continuous space. It has been used to solve a range of optimization problems, including neural network training and function minimization. Recently, it was successful adapted to optimization in binary spaces, presenting good performance also when applied to discontinuous objective functions and used in the optimization of many nonlinear functions and in artificial neural networks training. Engelbrecht and Merwe also explored the applicability of PSO to cluster data vector, by modifying its basic algorithm [Engelbrecht and Merwe, 2003].

Chemical database is designed to store chemical information, such as structure diagrams. Traditional chemical structure diagrams have been used to support various tasks in chemical research and development. Large chemical databases are expected to handle the storage and searching of information on millions of molecules taking terabytes of physical memory. An important feature in a chemical database system is the ability to quantify the degree of structural similarity between pairs, or larger groups, of molecules.

1.2 Problem Background

The development process of new drugs is a lengthy and costly procedure. The historical method of drug discovery is by trial-and-error testing of chemical substances on animals, and matching the apparent effects to treatments. The new method of drug design begins with knowledge of specific chemical responses in the body or target organism, and tailoring combinations of these to fit a treatment profile.

The process needs clustering process in order to choose compounds from each cluster representative of the structural content of the original compound database, classify substitute properties that are present in a dataset and summarize the classes of compounds that exist in a given dataset. The clustering process also can be used to view range of structural classes that contains a user-defined sub-structure, Analyze structure-activity relationship, and also predict unknown properties of compounds from other compounds in the same cluster.

There are challenges caused by large chemical space describing potential new drugs without side-effects, to find drug-like compounds from a database of thousands and millions of compounds. According to the *similar property principle*, structurally similar molecules will exhibit similar physiochemical and biological properties [Fink, November 1996].

Recently, several chemical databases that contain thousands or millions of chemical compound data have been developed. Based on that database, several grouping or clustering techniques developed to accelerate drug design processes.

The thousands or millions of chemical compound grouped based on their attributes also called descriptors. However, clustering is a difficult problem combinatorially [Jain, 1999].

1.3 Problem Statement

Based on the background given in previous section, looking for new technique of clustering of chemical compound data is very importance. The compound chemical data need to be clustered (grouped) into many cluster because some need especially in food and drug design. There are many methods and techniques which we are going to use could help us in best way to do that job. This study tries to applying Particle Swarm Optimization (PSO) to cluster chemical compound data. This study also observes about the performance PSO algorithm to clustering continuous and binary data. The study expect that PSO algorithm perform better than other algorithm because synergizing more than process into best result, also PSO had been known as good algorithm in term to search optimal solution through the search space.

1.4 Objectives

The objectives of this study are:

- To cluster chemical compound data using Particle Swarm Optimization (PSO) for both continuous and binary representation of chemical data.
- To utilize PSO in optimizing the clustering results produced by other clustering algorithm on chemical data.
- To analyze performance of PSO algorithm by comparing with Ward's algorithm in clustering different representation chemical compound data; continuous and binary.

1.5 Scope of Study

In order to achieve the objective stated above, the scopes of this study are limited as follows:

1. The clustering algorithm that will be used is Particle Swarm Optimization based algorithm.
2. Particle Swarm Optimization will also be applied to optimize the results of K-means clustering.
3. Ward's algorithm also implemented as comparison.
4. Data to be clustered are the chemical compound data from MDDR database.
5. The representations of data are continuous and binary data format.

1.6 Significance of Study

After the performance of PSO based clustering is analyzed, we can determine how effective this clustering techniques. This is important to identify the suitable technique for chemical database clustering and can be implemented in real world application.

1.7 Organization of Report

This report is mainly divided into five chapters. The first chapter provides an introduction and brief overview of the project including the problem background, problem statement, objective, scope, and contribution of the study. Chapter 2 reviews the literature background on previous study on data clustering. Chapter 3 review the literature background on previous studies on clustering of chemical compounds. This includes the results roles and constraints controlling the process of clustering these compounds as well as the methods used in clustering process. Chapter 4 covers the methodology of the research, which focuses on the application of the Particle Swarm Optimization (PSO) algorithm on clustering the chemical compounds. Chapter 5 presents the results obtained from the previously analyzed data. Chapter 6 contains the conclusions and recommendations of this project.